

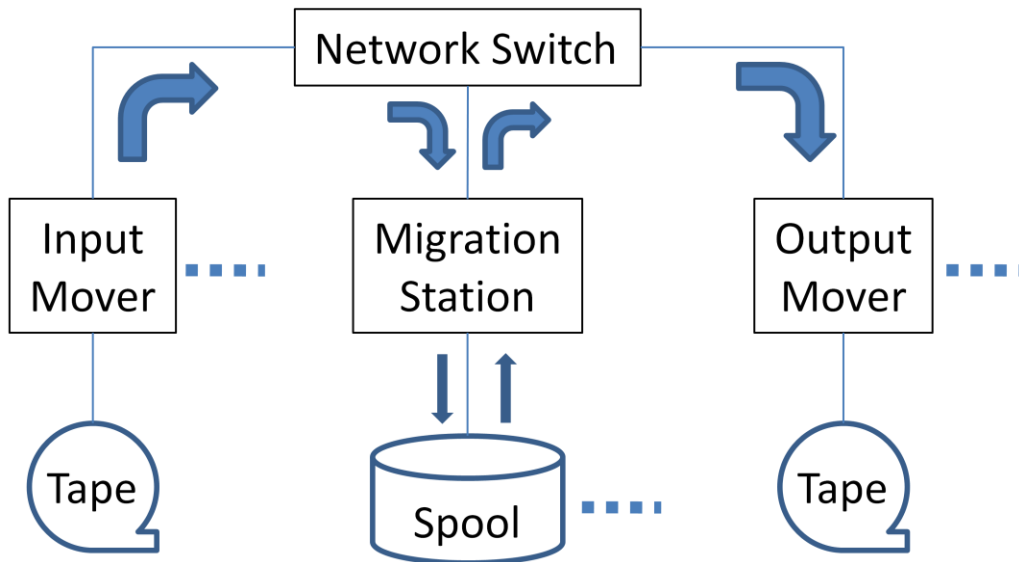
Comments on the state of media migration

Oct. 22, 2008

S. Fuess

Architecture:

A very rough schematic of the migration stand architecture is:



Current operations:

- 8 migration stations configured
- Running LTO2 to LTO4 migration

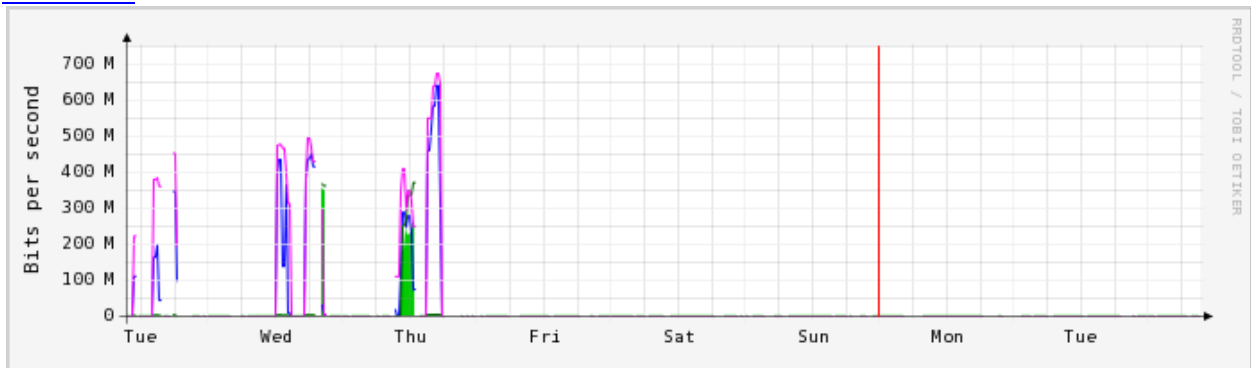
Analysis methodology:

- Checked network MRTG plots on migration servers
- Ran vmstat and iostat on migration servers to check disk rates

Network Observations:

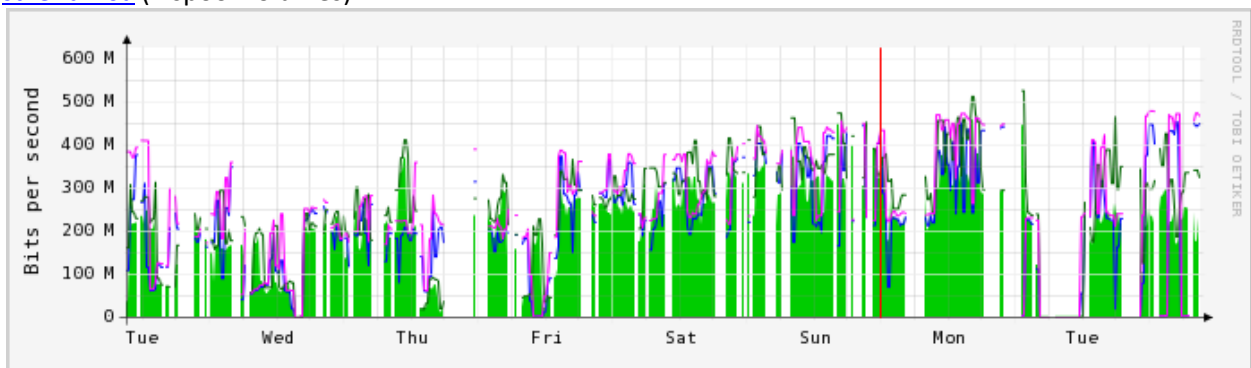
- Shown below are the MRTG network plots for each of the 8 migration stations for the previous week.
- Each station has a bonded dual-Gb connection to switch r-s-fcc2-server.
- In the following:
 - Plots show 30-minute average with 5-minute peak
 - Green = “in” to switch, thus “out” of migration station to LTO4 mover
 - Dark Green = 5-minute peak of Green
 - Blue = “out” from switch, thus “in” to migration server from LTO2 mover
 - Magenta = 5-minute peak of Blue

[stkendm5a](#)



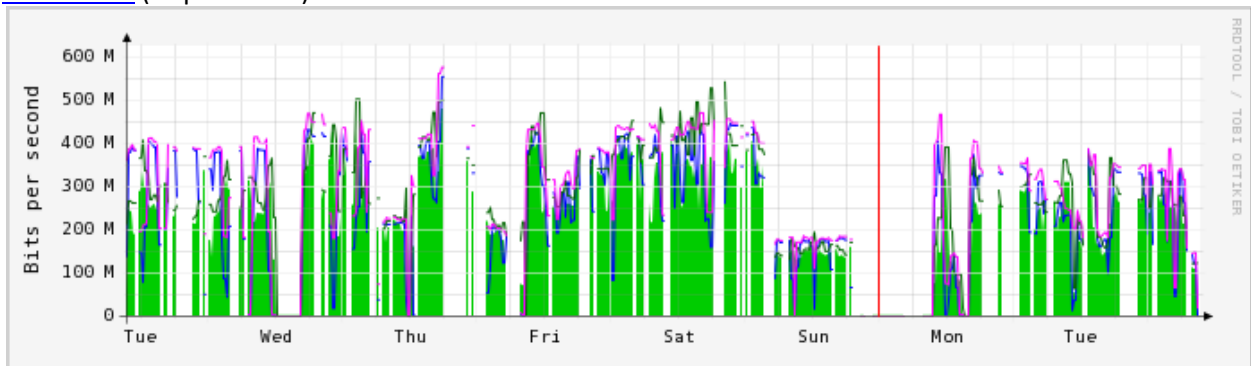
- This migration station is mostly idle.

[stkendm6a](#) (2 spool volumes)



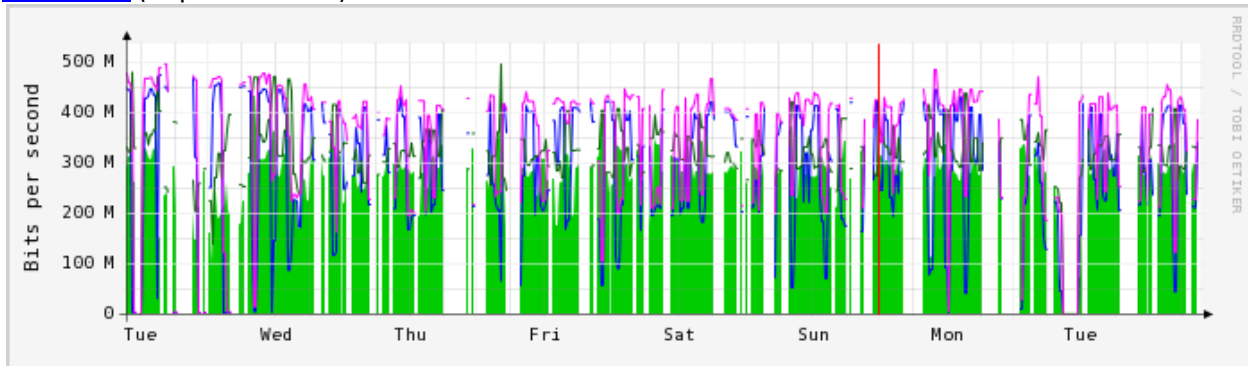
- While active, 30-minute average to station ranges from 200-450 Mb/s (25-56 MB/s).
- While active, 30-minute average from station ranges from 200-350 Mb/s (25-44 MB/s).
- Duty factor for this station is ~60%.

[stkendm7a](#) (4 spool disks)



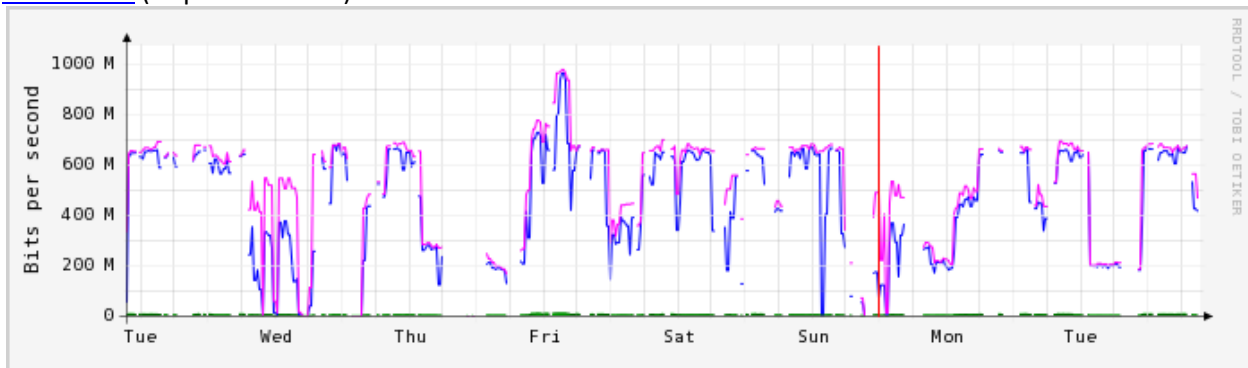
- While active, 30-minute average to station ranges from 200-450 Mb/s (25-56 MB/s).
- While active, 30-minute average from station ranges from 200-400 Mb/s (25-50 MB/s).
- Duty factor for this station is ~60%.

[stkensm8a](#) (2 spool volumes)



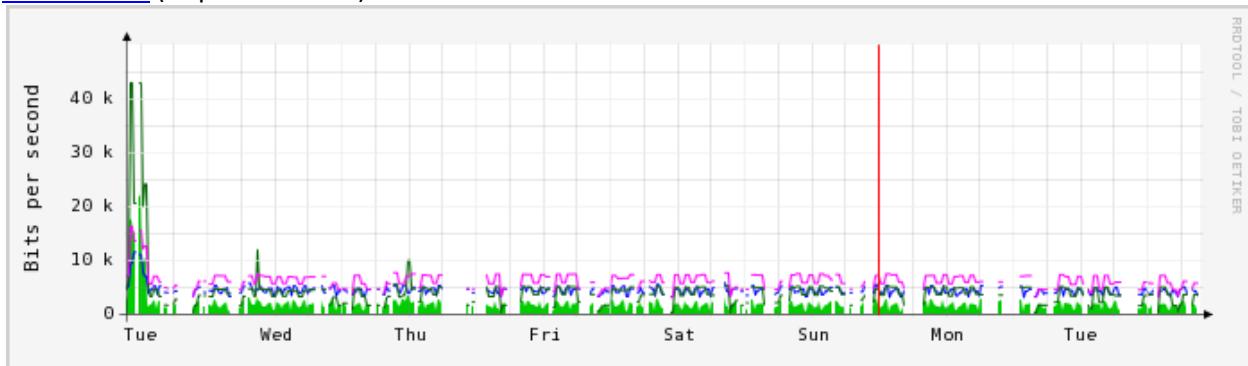
- While active, 30-minute average to station is ~450 Mb/s (56 MB/s).
- While active, 30-minute average from station ranges from 250-300 Mb/s (31-37 MB/s).
- Duty factor for this station is ~80%.

[stkendm9a](#) (2 spool volumes)



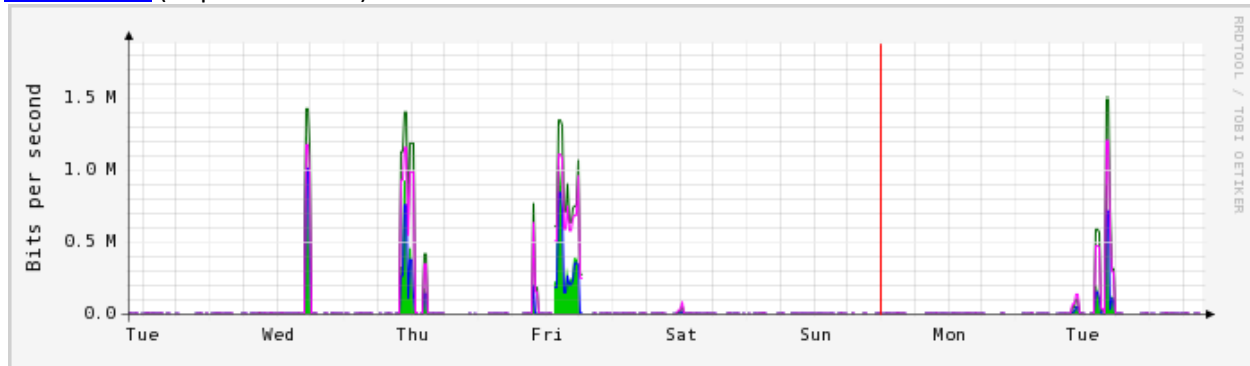
- This station is likely used for scanning of migrated tape, reading from LTO4 to station at rates of 600-900 Mb/s (75-112 MB/s).

[stkendm10a](#) (2 spool volumes)



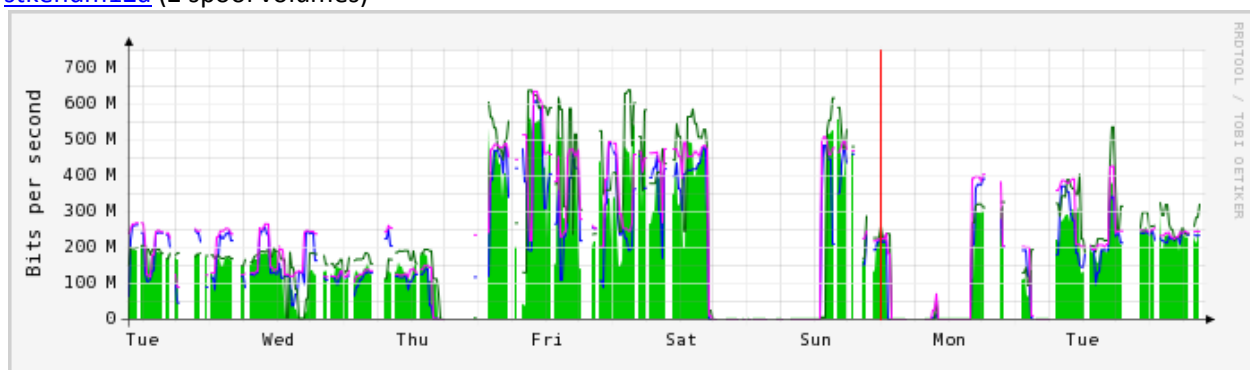
- This station is mostly idle.

[stkendm11a](#) (2 spool volumes)



- This station is mostly idle.

[stkendm12a](#) (2 spool volumes)



- While active, 30-minute average to station ranges from 200-450 Mb/s (25-56 MB/s).
- While active, 30-minute average from station ranges from 200-400 Mb/s (25-50 MB/s).
- Duty factor for this station is ~40%.

Disk Rate Observations:

Disk I/O rates were observed on idle (to run tests) and active migration stations:

- Tests were performed on stkendm10a while it was otherwise idle. This node has 8GB of memory and one quad-core 2.86 GHz Intel X3360 processor. Two disk arrays are configured from a (shared with other stkendm nodes) SataBeast – each with 535 GB total space .

Note: this is a question, as the thinking was that each volume was constructed as a RAID-1 stripe set of five 1 TB spindles – and hence would have expected 5000 GB available space.

One or more read or write instance of dd (block sizes of 1024 and 131072 were tried with little difference) was run. Observations (from vmstat) are:

- A single write to disk averages 175 MB/s. During the write operations the effect of the buffer cache is noted, where at the 5s buffer flush interval the rate to disk peaks near 270 MB/s. It is not clear why this peak rate can't be sustained.
- A single read from disk averages 125 MB/s (small block size) to 200 MB/s (large block size).
- A test with a simultaneous read and write to the same volume showed a write rate of ~175 MB/s and a read rate near zero until the write completes, at which time the read proceeds at ~170 MB/s (large block size). This behavior of the write dominating the read is a known feature of the Linux "cfq" elevator (I/O scheduler). In practice it is

unlikely that any write operation will saturate the scheduler, and hence intervening read operations should proceed.

- A test with two simultaneous write operations to distinct volumes gives an aggregate write rate of between 200 and 340 MB/s. The variation in instantaneous rates is likely due to buffer caching.
- A test with a simultaneous write operation to one volume and a read operation from a different volume yields a write rate of 100-150 MB/s and a read rate of ~122 MB/s. At these rates the idle time of the processor is near zero, implying that there are insufficient cycles to handle any higher rate.
- Disk rates were measured (with vmstat or iostat) during normal migration operations, which consisted of a single tape being read and a single tape being written. Observed rates were:
 - The **rate to disk** ("bo" in vmstat) typically shows a pattern of ~100 MB/s for one second, followed by two seconds of no output – giving an **average of ~33 MB/s (260 Mb/s)**. This rate is consistent with what is expected from reading the LTO2 drive. This pattern is repeated for the duration of reading a single file, type 100 MB to 1 GB in size. Between files the rate drops to zero.

Note that the standard pdflush dirty cache flushing period is 30 seconds, and pdflush wakes every 5 seconds to check. With the observed 3 second period, the hypothesis is that in this period greater than 10% of the buffer cache becomes dirty; hence the 100 MB written at this time implies a 1 GB buffer cache is being employed. (This could be easily studied further.)

The longer term average rate from input tape to disk buffer is less than 33 MB/s, as reduced by the null periods between file transfers. The network MRTG plots illustrate the rates averaged over longer periods.

- The **rate from disk** ("bi" in vmstat) is **consistently 75 MB/s (600 Mb/s)** during the period a file is being transferred. This is consistent with an LTO4 write speed of 60-120 MB/s. In between these transfers the rate drops to zero. The idle time is due to file metadata operations; it is expected to be ~2 secs, but was observed to be ~10 secs for many of the transfers studied – a possible indication of a problem.

The longer term average rate from disk buffer to output tape is less than 75 MB/s, as reduced by the null periods between file transfers. The network MRTG plots illustrate the rates averaged over longer periods.

Conclusions:

- The rate from LTO2 to spool disk averages 33 MB/s (260 Mb/s) while a single file transfer is active. The network plots average a minimum of 5 minutes, so display this active rate averaged with the inactive between-file periods. The network rates show 200-450 Mb/s, implying an efficiency of $200/260 = 77\%$ for a single input stream, or $450/520 = 87\%$ for dual input streams.
- The rate from spool disk to LTO4 is 75 MB/s (600 Mb/s) during the period a file is being transferred. The network plots average a minimum of 5 minutes, so display this active rate averaged with the inactive period between file transfers. The network rates show 200-400 Mb/s, implying an efficiency of $200/600 = 33\%$ to $400/600 = 66\%$. See the next bullet for an interpretation of these ranges.
- For a single instance of the migration application running on a migration station, the architecture allows for up to two input streams (LTO2 tapes) being read and spooled to separate buffer disk volumes, and up to two output streams (LTO4 tapes) being written, where each stream is associated with a single file family. The possibility of multiple streams explains what is observed in the network plots: two input streams, each maximally 33 MB/s, with the

inefficiencies associated with file activity included produce the observed 5-minute average of ~450 Mb/s. The interpretation of the output streams is more complicated: the output is essentially starved for data, hence any 5-minute or longer average is unlikely to show elevated rates.

Recommendations:

- At least during the 1-week period studied, the migration stations are not being kept active. Using an eyeball estimate of the efficiency of operation by noting the “duty factor” seen in the network plots, the station utilization is:

Station	Duty Factor
dm5	0.0
dm6	0.6
dm7	0.6
dm8	0.8
dm9	(scanning)
dm10	0.0
dm11	0.0
dm12	0.4
Total	2.4

The existing migration station utilization should be easily improved by keeping more jobs queued.

It is possible to perform a rough calculation for overall throughput using values observed during this study. An active single input stream reads tape at 33 MB/s. Using the net duty factor of 2.4 from above, this gives an aggregate rate of $2.4 * 33 \text{ MB/s} = 80 \text{ MB/s}$, equivalent to 12.5 seconds per GB. Thus 1 PB of data requires 4.8 months to migrate.

- The process seems to be starved for data, with the input from the (possibly multiple) read of the LTO2 and spooling to buffer disk not supplying data at a rate to keep the output stream to the LTO4 busy. There is sufficient pool disk space to buffer a considerably larger amount of data. We should consider increasing the ratio of input to output drives from 2:1 to 3:1.
- If the process is no longer starved for input data, then the slowest point may become the process of writing to the output stream. In this case, there is a need to reduce the inactive period between file writes to a minimum: the often observed ~10 seconds per file is too long.
- There is no evidence that running a single migration instance on a migration station is close to saturating any node, disk, or network resource. We should test running two instances on each station. This may push the limit of disk I/O capability with greater read/write contention; this problem could be rectified with greater or better organized buffer disk resources. Four simultaneous reads (2 per instance) at 33 MB/s each yields ~1 Gb/s into the station, which is within the capability of the bonded pair of Gb NICs. Four simultaneous writes (2 per instance, but unlikely) of 75 MB/s each totals 2.4 Gb/s, exceeding the network capacity; if this situation was encountered then output would be restricted.